IMAGE RESTORATION USING A STOCHASTIC VARIANT OF THE ALTERNATING DIRECTION METHOD OF MULTIPLIERS

Shunsuke Ono[†], *Masao Yamagishi*[†], *Takamichi Miyata*[††], *and Itsuo Kumazawa*[†]

[†]Tokyo Institute of Technology, [††]Chiba Institute of Technology

ABSTRACT

We propose an efficient image restoration framework based on stochastic optimization. Image restoration usually requires some iterative methods for solving optimization problems that characterize restored images, where the multiplication of the observation matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ and variables has to be computed at each iteration. If an efficient implementation of the multiplication (e.g., using FFT) is unavailable, its computational cost becomes $\mathcal{O}(MN)$, which is quite expensive since both N and M are usually large in image restoration. Our method needs to load and apply only a part of the observation matrix of size $\frac{M}{b} \times N$ (b: the number of parts), so that the computational cost is only $\mathcal{O}(\frac{MN}{b})$. Moreover, the proposed method accepts various nonsmooth objectives effective for image restoration. Experiments on compressed sensing reconstruction and non-uniform deblurring show the advantage of the proposed method over state-of-the-art proximal optimization methods.

Index Terms— Image restoration, stochastic optimization

1. INTRODUCTION

Image restoration, such as deblurring and compressed sensing (CS) reconstruction, is a fundamental problem in image processing. Most image restoration problems can be seen as inverse problems of the form: $\mathbf{v} = \mathcal{D}(\mathbf{\Phi}\bar{\mathbf{u}})$, where $\bar{\mathbf{u}} \in \mathbb{R}^N$ is an unknown original image of interest, $\mathbf{v} \in \mathbb{R}^M$ is an observation vector, $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is a matrix representing an observation process (e.g., blur), and $\mathcal{D} : \mathbb{R}^M \to \mathbb{R}^M$ is a noise contamination process that is not necessarily additive.

Variational approaches using nonsmooth regularization, e.g., total variation (TV) [1], have already been proven to be effective for image restoration. This has led to the demand for efficient algorithms to solve large-scale (usually $M, N > 10^4$) nonsmooth optimization problems. A successful class of such algorithms is *first-order proximal optimization methods* [2]. In particular, linearized variants of the alternating direction method of multipliers (L-ADMM) [3, 4] and the primal-dual splitting methods (PDS) [5, 6, 7] are preferable in the sense that they do not require matrix inversion.

However, an important issue still remains: at every iteration, proximal optimization methods have to compute the multiplication of the observation matrix $\mathbf{\Phi}$ and variables. For denoising and inpainting, this does not matter since $\mathbf{\Phi}$ is a simple diagonal matrix. For uniform deblurring with appropriate boundary conditions, the multiplication can be efficiently computed via FFT. On the other hand, such an efficient computation is unavailable in the case of non-uniform deblurring due to spatially-varying blur kernels, so that existing non-uniform deblurring methods employ locally-uniform kernel approximation or focus on specific blur types [8, 9, 10]. Similarly, for CS reconstruction with random measurements [11, 12], each entry of $\mathbf{\Phi}$ is a sample of random variables (e.g., Gaussian), implying that $\mathbf{\Phi}$ has no specific structure allowing efficient implementation of the multiplication. For such *complicated* $\mathbf{\Phi}$, proximal optimization methods require $\mathcal{O}(MN)$ computation per iteration, which is expensive since both M and N are large.

To overcome the difficulty, this paper proposes an efficient image restoration framework based on a recently proposed stochastic proximal optimization method: *Stochastic Dual Coordinate Ascent with ADMM* (SDCA-ADMM) [13]. To the best of our knowledge, this work is the first attempt to leverage stochastic proximal optimization to resolve image restoration with complicated $\mathbf{\Phi}$ (see Remark 1 for related work).[1] In our framework, the observation matrix $\mathbf{\Phi}$ is decomposed into b sub-matrices $\mathbf{\Phi}_{\mathcal{I}_k} \in \mathbb{R}^{\frac{M}{b} \times N}$ ($k = 1, \ldots, b$), where $\mathcal{I}_k$ is the kth *mini-batch* containing the indices of the rows of $\mathbf{\Phi}$ that construct $\mathbf{\Phi}_{\mathcal{I}_k}$. Then, in optimization, only a randomly chosen $\mathbf{\Phi}_{\mathcal{I}_k}$ is activated per iteration, i.e., the computational cost is $\mathcal{O}(\frac{MN}{b})$. Hence, the proposed method is much more efficient than non-stochastic proximal optimization methods in image restoration with complicated $\mathbf{\Phi}$, demonstrated by our experiments.

2. SDCA-ADMM

In the machine learning literature, the stochastic dual coordinate ascent with ADMM (SDCA-ADMM) [13] was proposed to solve:

$$\min_{\mathbf{w} \in \mathbb{R}^N} \frac{1}{M} \sum_{m=1}^{M} f_m(\mathbf{z}_m^\top \mathbf{w}) + \psi(\mathbf{B}^\top \mathbf{w}), \tag{1}$$

where $\mathbf{w}$ is the weight vector one wants to learn, $\mathbf{z}_1, \ldots, \mathbf{z}_M \in \mathbb{R}^N$ are given vectors, $f_m : \mathbb{R} \to (-\infty, \infty]$ is a loss function for the mth sample, and $\psi \circ \mathbf{B}^\top : \mathbb{R}^N \to (-\infty, \infty]$ is a regularization function ($\mathbf{B} \in \mathbb{R}^{N \times K}$, $\psi : \mathbb{R}^K \to (-\infty, \infty]$). Assume the following: f_m and ψ are proper lower semicontinuous convex, and their *proximity operators* [15] are easy to compute, where the proximity operator of index $\gamma > 0$ of $g \in \Gamma_0(\mathbb{R}^N)$[2] is defined by

$$\mathrm{prox}_{\gamma g} : \mathbb{R}^N \to \mathbb{R}^N : \mathbf{x} \mapsto \operatorname*{argmin}_{\mathbf{y}} g(\mathbf{y}) + \tfrac{1}{2\gamma}\|\mathbf{y} - \mathbf{x}\|^2.$$

Then, SDCA-ADMM solves Prob. (1) via its dual problem:

$$\min_{\mathbf{x} \in \mathbb{R}^M, \mathbf{y} \in \mathbb{R}^K} \frac{1}{M} \sum_{m=1}^{M} f_m^*(x_m) + \psi^*(\tfrac{1}{M}\mathbf{y}) \text{ s.t. } \mathbf{Z}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{0},$$

where $\mathbf{Z} := (\mathbf{z}_1 \cdots \mathbf{z}_M) \in \mathbb{R}^{N \times M}$ ($\cdot^*$ means convex conjugation).

Let $\mathcal{I}_k \subset \{1, \ldots, M\}$ be the kth mini-batch including the indices of a subset of samples ($k = 1, \ldots, b$, and b is the number of mini-batches). All the mini-batches satisfy $\cup_{k=1}^{b} \mathcal{I}_k = \{1, \ldots, M\}$

The work was supported by JSPS Grants-in-Aid: 15H06197; 15K06078.

[1] The preliminary version of the work appeared in a technical report [14].

[2] The set of all proper lower semicontinuous convex functions on $\mathbb{R}^N$ is denoted by $\Gamma_0(\mathbb{R}^N)$.

Algorithm 1: SDCA-ADMM

Choose $\mathbf{x}^{(0)}, \mathbf{y}^{(0)}, \mathbf{w}^{(0)}, \boldsymbol{\xi}^{(0)}$, $\rho > 0$, $\tau > 0$, $\eta_{\mathbf{B}} > (\sigma_1(\mathbf{B}))^2$, $\eta_{\mathbf{Z}_{\mathcal{I}_k}} > (\sigma_1(\mathbf{Z}_{\mathcal{I}_k}))^2$, and set $n = 1$.

while *A stopping criterion is not satisfied* **do**

Choose $k \in \{1, \dots, b\}$ uniformly at random.

1 $\mathbf{y}^{(n)} \leftarrow \operatorname{prox}_{\frac{M}{\rho\eta_{\mathbf{B}}}\psi^*}(\frac{1}{M}(\mathbf{y}^{(n-1)} + \frac{1}{\rho\eta_{\mathbf{B}}}\mathbf{B}^\top(\mathbf{w}^{(n-1)} - \rho(\boldsymbol{\xi}^{(n-1)} + \mathbf{B}\mathbf{y}^{(n-1)}))))$

2 $\mathbf{x}^{(n)}_{\mathcal{I}_k} \leftarrow \operatorname{prox}_{\frac{1}{\rho\eta_{\mathbf{Z}_{\mathcal{I}_k}}} f^*_{\mathcal{I}_k}}(\mathbf{x}^{(n-1)}_{\mathcal{I}_k} + \frac{1}{\rho\eta_{\mathbf{Z}_{\mathcal{I}_k}}}\mathbf{Z}^\top_{\mathcal{I}_k}(\mathbf{w}^{(n-1)} - \rho(\boldsymbol{\xi}^{(n-1)} + \mathbf{B}\mathbf{y}^{(n)})))$

3 $\boldsymbol{\xi}^{(n)} \leftarrow \boldsymbol{\xi}^{(n-1)} + \mathbf{Z}_{\mathcal{I}_k}(\mathbf{x}^{(n)}_{\mathcal{I}_k} - \mathbf{x}^{(n-1)}_{\mathcal{I}_k})$

4 $\mathbf{w}^{(n)} \leftarrow \mathbf{w}^{(n-1)} - \tau\rho(M(\boldsymbol{\xi}^{(n)} + \mathbf{B}\mathbf{y}^{(n)}) - (M - \frac{M}{b})(\boldsymbol{\xi}^{(n-1)} + \mathbf{B}\mathbf{y}^{(n-1)}))$

$n \leftarrow n + 1$

and $\mathcal{I}_k \cap \mathcal{I}_{k' \neq k} = \emptyset$. Then, at each iteration, SDCA-ADMM randomly chooses one with probability $1/b$ from all the mini-batches, and updates variables by only using the samples w.r.t. the mini-batch. The detailed computation of SDCA-ADMM is given in Alg. 1, where $f_{\mathcal{I}_k}(\mathbf{x}_{\mathcal{I}_k}) := \sum_{m \in \mathcal{I}_k} f_m(x_m)$, and $\mathbf{Z}_{\mathcal{I}_k} \in \mathbb{R}^{N \times \frac{M}{b}}$ and $\mathbf{x}_{\mathcal{I}_k} \in \mathbb{R}^{\frac{M}{b}}$ are the submatrix of $\mathbf{Z}$ and the subvector of $\mathbf{x}$ w.r.t. the kth mini-batch, respectively ($\sigma_1(\cdot)$ is the largest singular value of $(\cdot)$). Note that using the proximity operator of $g \in \Gamma_0(\mathbb{R}^N)$, the proximity operator of $g^* \in \Gamma_0(\mathbb{R}^N)$ can be expressed as $\operatorname{prox}_{\gamma g^*}(\mathbf{x}) = \mathbf{x} - \gamma \operatorname{prox}_{\frac{1}{\gamma} g}(\frac{1}{\gamma}\mathbf{x})$ [16, Theorem 14.3(ii)].

Remark 1 (Other stochastic proximal optimization methods). Although we adopt SDCA-ADMM in our framework due to its formulation, there are several other stochastic proximal optimization methods that can be applied to (1) with more specific structures. The first one is the SAGA algorithm [17], which can solve (1) when f_m is Lipschitz-differentiable and $\operatorname{prox}_{\psi \circ \mathbf{B}^\top}$ is computable. If f_m and ψ can be decomposed w.r.t. sub-vectors of $\mathbf{w}$, the stochastic primal-dual proximal algorithm [18, 19] would be another choice for solving (1). We refer the readers to [20] for more information.

3. PROPOSED METHOD

3.1. Problem formulation

We consider the following variational image restoration:

$$\min_{\mathbf{u} \in \mathbb{R}^N} \mathcal{F}_{\mathbf{v}}(\mathbf{\Phi}\mathbf{u}) + \mathcal{R}(\mathbf{F}\mathbf{u}) \ \text{ s.t. } \mathbf{u} \in C, \tag{2}$$

where $\mathcal{F}_{\mathbf{v}} \in \Gamma_0(\mathbb{R}^M)$ is a data-fidelity function, $\mathcal{R} \circ \mathbf{F} : \mathbb{R}^N \to (-\infty, \infty]$ is a regularization function ($\mathbf{F} \in \mathbb{R}^{L \times N}$, $\mathcal{R} \in \Gamma_0(\mathbb{R}^L)$), and $C \subset \mathbb{R}^N$ is a closed convex constraint on $\mathbf{u}$. We assume
(a1) $\mathcal{F}_{\mathbf{v}}$ is *separable*, i.e., $\mathcal{F}_{\mathbf{v}}(\mathbf{x}) = \sum_{m=1}^M \mathcal{F}_m(x_m)$.
(a2) The computational costs of $\operatorname{prox}_{\gamma \mathcal{F}_m}$ and $\operatorname{prox}_{\gamma \mathcal{R}}$ are $\mathcal{O}(1)$ and $\mathcal{O}(L)$, respectively.
(a3) The multiplication of $\mathbf{F}$ (and $\mathbf{F}^\top$) is $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$.
(a4) The computational cost of P_C is $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$.[3]

Remark 2 (Examples of $\mathcal{F}_{\mathbf{v}}$, $\mathcal{R} \circ \mathbf{F}$, and C).
(Data-fidelity function $\mathcal{F}_{\mathbf{v}}$) The ℓ_2 norm would be the most popular one and clearly separable, given by $\mathcal{F}_{\mathbf{v}}(\mathbf{x}) := \frac{\mu}{2}\|\mathbf{x} - \mathbf{v}\|^2 = \frac{\mu}{2}\sum_{m=1}^M (x_m - v_m)^2$. The ℓ_1 norm is also useful for data-fidelity measure, especially in the case where $\mathbf{v}$ contains outliers. It is defined by $\mathcal{F}_{\mathbf{v}}(\mathbf{x}) := \mu\|\mathbf{x} - \mathbf{v}\|_1 = \mu\sum_{m=1}^M |x_m - v_m|$, i.e., separable. In the case of Poisson noise contamination, the *generalized Kulback-Leibler divergence*, which is also separable, is known as a suitable data-fidelity function (the definition can be found in [21]). The proximity operators of the above examples satisfy (a2).
(Regularization function $\mathcal{R} \circ \mathbf{F}$) TV [1] and its vectorial variants, e.g., [22, 23, 24], are well-known edge-preserving regularizers for images. In this case, $\mathcal{R}$ is a norm, usually the mixed $\ell_{1,2}$ norm, and $\mathbf{F}$ is a discrete gradient operator. The proximity operator of the mixed $\ell_{1,2}$ norm is available with $\mathcal{O}(N)$, and the computation of the discrete gradient operator is also $\mathcal{O}(N)$. Another well-known example is frame regularization relying on the sparsity of images in some transformed domain. In this case, $\mathcal{R}$ is the ℓ_1 norm, whose proximity operator is computable with $\mathcal{O}(L)$ (L is the number of the frame coefficients), and $\mathbf{F}$ is a frame analysis operator, e.g., wavelet and curvelet [25]. Most well-designed frame analysis operations can be performed in $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$. Nonlocal regularization [26, 27, 28] and regularizaiton using learned operators [29, 30] can also be considered in this framework if a nonlocal/learned analysis operator $\mathbf{F}$ allowing efficient implementation.
(Constraint C) One can impose some additional knowledge on the original image $\bar{\mathbf{u}}$. A simple example is a box constraint that represents a known dynamic range, e.g., $C := [0, 255]^N$ for eight-bit images. Imposing this type of bounded closed convex constraints also guarantees the existence of the minimizer of (2).

3.2. Reformulation, mini-batch construction, and algorithm

By noting the separability of $\mathcal{F}_{\mathbf{v}}$ and by using the indicator function[4] of C, Prob. (2) can be rewritten as

$$\min_{\mathbf{u} \in \mathbb{R}^N} \textstyle\sum_{m=1}^M \mathcal{F}_m(\boldsymbol{\phi}_m^\top \mathbf{u}) + \mathcal{R}(\mathbf{F}\mathbf{u}) + \iota_C(\mathbf{u}), \tag{3}$$

where $\boldsymbol{\phi}_m \in \mathbb{R}^N$ is the mth row vector of $\mathbf{\Phi}$, i.e., $\mathbf{\Phi}^\top = (\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_M)$. Let us define $\mathbf{B} := (\mathbf{F}^\top \ \mathbf{I}) \in \mathbb{R}^{N \times (L+N)}$ and $\psi : \mathbb{R}^{L+N} \to (-\infty, \infty] : \mathbf{y} \mapsto \mathcal{R}(\mathbf{y}_L) + \iota_C(\mathbf{y}_N)$, where $\mathbf{y} = (\mathbf{y}_L^\top \ \mathbf{y}_N^\top)^\top$. Then, Prob. (3) can be reformulated into

$$\min_{\mathbf{u} \in \mathbb{R}^N} \textstyle\sum_{m=1}^M \mathcal{F}_m(\boldsymbol{\phi}_m^\top \mathbf{u}) + \psi(\mathbf{B}^\top \mathbf{u}), \tag{4}$$

which is equivalent to Prob. (1) (except the constant $\frac{1}{M}$). Finally, as in the proof of [13, Lemma 1], using Fenchel-Rockafellar duality [16, Definition 15.19], the dual problem of (4) is obtained as

$$\min_{\mathbf{x} \in \mathbb{R}^M, \mathbf{y} \in \mathbb{R}^{L+N}} \textstyle\sum_{m=1}^M \mathcal{F}_m^*(x_m) + \psi^*(\mathbf{y}) \ \text{ s.t. } \mathbf{\Phi}^\top \mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{0}. \tag{5}$$

When we apply SDCA-ADMM to Prob. (5), constructing mini-batches suitable for the structure of the problem is quite important for fast convergence. Indeed, according to the analysis of SDCA-ADMM in [13], the convergence rate of SDCA-ADMM becomes worse when samples in a mini-batch are strongly correlated to each other. Since each entry of the observation vector $\mathbf{v}$ in image restoration corresponds to each sample in machine learning, this phenomenon should be carefully considered in the proposed method. Indeed, the *spatial correlation* of pixels is usually very strong, and this correlation would be propagated to entries of the

[3] Given a nonempty closed convex set $C \subset \mathbb{R}^N$, the *projection onto* C is defined by $P_C : \mathbb{R}^N \to \mathbb{R}^N : \mathbf{x} \mapsto \operatorname*{argmin}_{\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|_2$.

[4] For a given nonempty closed convex set $C \in \mathbb{R}^N$, the indicator function of C is defined by $\iota_C(\mathbf{x}) := 0$, if $\mathbf{x} \in C$; ∞, otherwise. The proximity operator of ι_C is equivalent to the projection onto C.

Algorithm 2: Solver for Prob. (2) based on SDCA-ADMM

Choose $\mathbf{x}^{(0)}, \mathbf{y}_L^{(0)}, \mathbf{y}_N^{(0)}, \mathbf{u}^{(0)}, \boldsymbol{\xi}^{(0)}, \mathbf{t}^{(0)}, \rho > 0, \tau > 0$, $\eta_{\mathbf{B}} > (\sigma_1(\mathbf{B}))^2$, $\eta_{\boldsymbol{\Phi}_{\mathcal{I}_k}} > (\sigma_1(\boldsymbol{\Phi}_{\mathcal{I}_k}))^2$, and set $n = 1$

while *A stopping criterion is not satisfied* **do**

Choose $k \in \{1, \ldots, b\}$ uniformly at random.

1 $\mathbf{r}^{(n)} = \mathbf{u}^{(n-1)} - \rho(\boldsymbol{\xi}^{(n-1)} + \mathbf{t}^{(n-1)})$

2 $\mathbf{q}_L^{(n)} = \mathbf{y}_L^{(n-1)} + \frac{1}{\rho\eta_{\mathbf{B}}}\mathbf{F}\mathbf{r}^{(n)}$

3 $\mathbf{q}_N^{(n)} = \mathbf{y}_N^{(n-1)} + \frac{1}{\rho\eta_{\mathbf{B}}}\mathbf{r}^{(n)}$

4 $\mathbf{y}_L^{(n)} \leftarrow \mathbf{q}_L^{(n)} - \frac{1}{\rho\eta_{\mathbf{B}}} \operatorname{prox}_{\rho\eta_{\mathbf{B}}\mathcal{R}}(\rho\eta_{\mathbf{B}}\mathbf{q}_L^{(n)})$

5 $\mathbf{y}_N^{(n)} \leftarrow \mathbf{q}_N^{(n)} - \frac{1}{\rho\eta_{\mathbf{B}}} P_C(\rho\eta_{\mathbf{B}}\mathbf{q}_N^{(n)})$

6 $\mathbf{t}^{(n)} = \mathbf{F}^\top \mathbf{y}_L^{(n)} + \mathbf{y}_N^{(n)}$

7 $\mathbf{s}^{(n)} = \mathbf{u}^{(n-1)} - \rho(\boldsymbol{\xi}^{(n-1)} + \mathbf{t}^{(n)})$

8 $\mathbf{p}_{\mathcal{I}_k}^{(n)} = \mathbf{x}_{\mathcal{I}_k}^{(n-1)} + \frac{1}{\rho\eta_{\boldsymbol{\Phi}_{\mathcal{I}_k}}}\boldsymbol{\Phi}_{\mathcal{I}_k}\mathbf{s}^{(n)}$

9 $\mathbf{x}_{\mathcal{I}_k}^{(n)} \leftarrow \mathbf{p}_{\mathcal{I}_k}^{(n)} - \frac{1}{\rho\eta_{\boldsymbol{\Phi}_{\mathcal{I}_k}}} \operatorname{prox}_{\rho\eta_{\mathbf{z}_{\mathcal{I}_k}}\mathcal{F}_{\mathcal{I}_k}}(\rho\eta_{\boldsymbol{\Phi}_{\mathcal{I}_k}}\mathbf{p}_{\mathcal{I}_k}^{(n)})$

10 $\boldsymbol{\xi}^{(n)} = \boldsymbol{\xi}^{(n-1)} + \boldsymbol{\Phi}_{\mathcal{I}_k}^\top(\mathbf{x}_{\mathcal{I}_k}^{(n)} - \mathbf{x}_{\mathcal{I}_k}^{(n-1)})$

11 $\mathbf{u}^{(n)} \leftarrow$
$\mathbf{u}^{(n-1)} - \tau\rho(M(\boldsymbol{\xi}^{(n)} + \mathbf{t}^{(n)}) - \frac{(b-1)M}{b}(\boldsymbol{\xi}^{(n-1)} + \mathbf{t}^{(n-1)}))$

$n \leftarrow n + 1$

observation vector $\mathbf{v}$ (depending on the structure of $\boldsymbol{\Phi}$). A typical case is deblurring, where the entries of $\mathbf{v}$ are blurred pixels.

To deal with such cases, we suggest to construct mini-batches via *spatially-uniform sampling*. Let $\mathbf{V} \in \mathbb{R}^{M_v \times M_h}$ be the spatially-correlated 2D form of $\mathbf{v}$ ($M_v M_h = M$). For simplicity, the number of mini-batches b is assumed to be a square number, and M_v and M_h are assumed to be divisible by $\sqrt{b}$. Then, entries of $\mathbf{V}$ belonging to the kth mini-batch are selected by the kth spatially-uniform sampling operator $T_k : \mathbb{R}^{M_v \times M_h} \to \mathbb{R}^{\frac{M_v}{\sqrt{b}} \times \frac{M_h}{\sqrt{b}}}$, where, for $p = 1, \ldots, \sqrt{b}$ and $q = 1, \ldots, \sqrt{b}$, set $k := q + (p-1)\sqrt{b}$ and

$$T_k(\mathbf{V}) = \begin{pmatrix} V_{p,q} & V_{p,q+\sqrt{b}} & \cdots & V_{p,M_h+q-\sqrt{b}} \\ V_{p+\sqrt{b},q} & V_{p+\sqrt{b},q+\sqrt{b}} & \cdots & V_{p+\sqrt{b},M_h+q-\sqrt{b}} \\ \vdots & \vdots & \ddots & \vdots \\ V_{M_v+p-\sqrt{b},q} & V_{M_v+p-\sqrt{b},q+\sqrt{b}} & \cdots & V_{M_v-p+\sqrt{b},M_h+q-\sqrt{b}} \end{pmatrix}.$$

Consequently, all the entries in one $T_k(\mathbf{V})$ are as far from each other as possible. Thus, by using this mini-batch construction strategy, we can alleviate the spatial correlation of the entries of each mini-batch.

Now we arrive at the point where we can apply SDCA-ADMM to solve Prob. (5), i.e., Prob. (2). Let $\boldsymbol{\Phi}_{\mathcal{I}_k} \in \mathbb{R}^{\frac{M}{b} \times N}$ be a submatrix of $\boldsymbol{\Phi}$ w.r.t. the kth mini-batch, and define $\mathcal{F}_{\mathcal{I}_k}(\mathbf{x}_{\mathcal{I}_k}) := \sum_{m \in \mathcal{I}_k} \mathcal{F}_m(x_m)$. The resulting algorithm is summarized in Alg. 2.

In Alg. 2, we can see how mini-batch construction affects the convergence behavior. Suppose that a subvector of $\mathbf{v}$ corresponding to a mini-batch has strong spatial correlation, i.e., every entry of the subvector is composed of a linear combination of the pixels in a local region. Then, the data-fidelity is evaluated only w.r.t. the region (step 9). On the other hand, the effect of the regularization is always global (step 4), so that in the other regions, the regularization is performed without considering data-fidelity, which would result in a slow convergence. Indeed, we will see in Sec. 4 that mini-batch construction significantly affects the convergence speed.

Remark 3 (Computational cost of Alg. 2). Alg. 2 only needs to compute $\boldsymbol{\Phi}_{\mathcal{I}_k}\mathbf{x}$ and $\boldsymbol{\Phi}_{\mathcal{I}_k}^\top \mathbf{y}$ once at each iteration, implying that the proposed method is much more efficient than existing non-stochastic proximal optimization methods that require the computations of $\boldsymbol{\Phi}\mathbf{x}$ and $\boldsymbol{\Phi}^\top\mathbf{y}$ at each iteration. We list the computational costs of the steps involving matrix application or proximal operation in Alg. 2.
Step 2 and 6: $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$ from (a4).
Step 4: $\mathcal{O}(L)$ from (a2).
Step 5: $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$ from (a4).
Step 8 and 10: $\mathcal{O}(\frac{MN}{b})$.
Step 9: $\mathcal{O}(\frac{M}{b})$ from (a1)-(a2).

Remark 4 (Convergence of Alg. 2). The convergence of SDCA-ADMM was analyzed under a strong convexity assumption [13], which implies that as of now, there is no convergence analysis for general convex objectives such as (5). However, both for stochastic and non-stochastic methods, such a strong convexity assumption is usually required to achieve a linear convergence rate but is not necessary to guarantee convergence (for example, the convergence of another stochastic variant of ADMM [31] was proved for general convex objectives). Indeed, Alg. 2 shows stable convergence in our experiments (see Sec. 4).

4. EXPERIMENTS

We examined the performance of the proposed method by comparing it with several state-of-the-art non-stochastic proximal optimization methods in two specific image restoration applications with complicated $\boldsymbol{\Phi}$: compressed sensing (CS) reconstruction and non-uniform deblurring, All experiments were performed using MATLAB (R2013a), on a Windows 8.1 laptop computer.

Methods for comparison. We compared the proposed method with the primal-dual splitting method (PDS) [5, 6] and the linearized alternating direction method of multipliers (L-ADMM) [4], which require no matrix inversion.

Design of Prob. (2). We employed (isotropic) TV [1] for grayscale images and its vectorial variant [22] for color images as the regularization function $\mathcal{R} \circ \mathbf{F}$ in Prob. (2). In this case, the matrix $\mathbf{F}$ is equal to $\mathbf{D} := (\mathbf{D}_v^\top \ \mathbf{D}_h^\top)^\top \in \mathbb{R}^{2N \times N}$, where $\mathbf{D}_v$ and $\mathbf{D}_h$ are the vertical and horizontal discrete gradient operators with Neumann boundary. Hence, $\mathbf{F}\mathbf{x}$ and $\mathbf{F}^\top\mathbf{y}$ can be computed with $\mathcal{O}(N)$ cost. The function $\mathcal{R}$ is the mixed $\ell_{1,2}$ norm defined by $\|\mathbf{x}\|_{1,2} := \sum_{i=1}^{|\mathcal{G}|} \sqrt{\sum_{j \in \mathcal{G}_i} x_j^2}$, where $\mathcal{G}_i$ is the index set including the indices of entries of $\mathbf{x}$ belonging to the ith group ($i = 1, \ldots, |\mathcal{G}|$). Specifically, one group consists of vertical and horizontal discrete gradients w.r.t. the ith pixel in the TV case. The proximity operator of $\|\cdot\|_{1,2}$ is given by a simple $\mathcal{O}(N)$ soft-thresholding type operation (see, e.g., [32])

For the data-fidelity function $\mathcal{F}_{\mathbf{v}}$, we used different ones in CS reconstruction and non-uniform deblurring (explained later).

For the constraint C, we imposed a dynamic range constraint $[0, 255]^N$, onto which the projection can be calculated by pushing the entries into $[0, 255]$, i.e., $\mathcal{O}(N)$ cost.

Parameter settings. For the proposed method, we employed the parameter settings suggested in [13], specifically, $\tau = \frac{1}{M}$, $\rho = 0.1$ and $\eta_{\mathbf{B}} = 1.1(\sigma_1(\mathbf{B}))^2$ in all the experiments. Since it is not realistic to use different $\eta_{\boldsymbol{\Phi}_{\mathcal{I}_k}}$ for each k, we fixed all of them to $1.1(\max_k \sigma_1(\boldsymbol{\Phi}_{\mathcal{I}_k}))^2$.

For PDS and L-ADMM, we adjusted their parameters that give best convergence behavior in each experiment, respectively.

Evaluation criterion. For evaluation of convergence, we define the normalized root mean square error (NRMSE) between the current estimate $\mathbf{u}^{(n)}$ and the optimal solution $\mathbf{u}^\star$ of Prob. (2), i.e., $\mathrm{NRMSE}_n := \|\mathbf{u}^{(n)} - \mathbf{u}^\star\| / \|\mathbf{u}^\star\|$. Since the optimal solution $\mathbf{u}^\star$ is

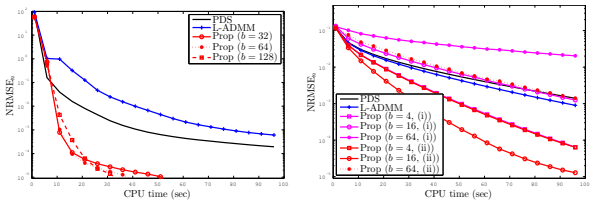

Fig. 1. Convergence profile of PDS, L-ADMM, and Alg. 2 (Prop) on CS reconstruction (left) and non-uniform deblurring (right).

analytically unavailable, it was pre-computed by PDS with 100000 iterations. For a fair comparison of stochastic and non-stochastic methods, the convergence curves of the proposed method were obtained after averaging uniformly 100 realizations.

4.1. Compressed sensing reconstruction

In CS reconstruction, we try to recover an original image $\bar{\mathbf{u}}$ from its incomplete measurements $\mathbf{v} = \mathbf{\Phi}\bar{\mathbf{u}}$, where $\mathbf{\Phi}$ is some measurement matrix of size $M \times N$ with $M < N$. Theoretically, employing random matrices as $\mathbf{\Phi}$ is in some sense an optimal strategy for a stable CS reconstruction [33, 34]. However, such random matrices are dense and have no specific structure allowing efficient implementation, so that the computations of $\mathbf{\Phi}\mathbf{x}$ and $\mathbf{\Phi}^\top\mathbf{y}$ become much expensive and memory inefficient in large-scale problems, as pointed out in [35]. The proposed method provides a resolution to this issue.

In this experiment, we solve

$$\min_{\mathbf{u}\in[0,255]^N} \|\mathbf{D}\mathbf{u}\|_{1,2} \ \text{s.t.}\ \mathbf{\Phi}\mathbf{u} = \mathbf{v}. \tag{6}$$

This problem appears different from Prob. (2) because the data-fidelity is expressed as the linear constraint $\mathbf{\Phi}\mathbf{u} = \mathbf{v}$, but by using M indicator functions ι_{E_m} ($m = 1, \ldots, M$) with $E_m := \{v_m\}$, Prob. (7) can be reduced to Prob. (2) as follows:

$$\min_{\mathbf{u}\in[0,255]^N} \textstyle\sum_{m=1}^M \iota_{E_m}(\boldsymbol{\phi}_m^\top\mathbf{u}) + \|\mathbf{D}\mathbf{u}\|_{1,2}.$$

Since E_m is a singleton, the computation of the proximity operator of ι_{E_m} (the projection onto E_m) is just replacing the input by v_m.

For a test image, we used a grayscale *Lena* image of size 128×128 ($N = 16384$). The measurement matrix $\mathbf{\Phi}$ was set to a 4096×16384 random Gaussian matrix ($M = \frac{N}{4}$), i.e., its entries are realizations of i.i.d. random variables from a Gaussian probability density function with mean zero and variance $\frac{1}{N}$. In this case, all the entries of $\mathbf{v}$ is sufficiently decorrelated to each other through the random measurement process, so that we can construct the mini-batches by simple partitioning of $\mathbf{v}$. Note that we use a relatively small image because PDS and L-ADMM have to load full $\mathbf{\Phi}$ at each iteration.

The left of Fig. 1 shows the convergence profile of PDS, L-ADMM, and Alg. 2 on the CS reconstruction experiment. For the proposed method, we tested the three different numbers of mini-batches: $b = 32, 64, 128$. One sees that the proposed method converges much faster than PDS and L-ADMM for all the numbers of mini-batches. The resulting images in Fig. 2 indicate the same PSNR (26.39 [dB]), which illustrates that Alg. 2 properly works.

4.2. Non-uniform deblurring

Non-uniform deblurring is a realistic but still challenging problem since the blur kernel is spatially variant, which precludes an efficient

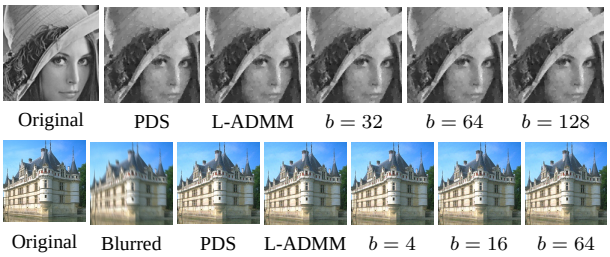

Fig. 2. Resulting images on CS reconstruction (top) and non-uniform deblurring (bottom).

optimization via FFT. In this experiment, a sharp image is restored from a blurred observation $\mathbf{v} = \mathbf{\Phi}\bar{\mathbf{u}} + \mathbf{n}$ by solving

$$\min_{\mathbf{u}\in[0,255]^N} \tfrac{\lambda}{2}\textstyle\sum_{m=1}^M(\boldsymbol{\phi}_m^\top\mathbf{u} - v_m)^2 + \|\mathbf{D}\mathbf{u}\|_{1,2}, \tag{7}$$

where $\mathbf{n}$ is an additive white Gaussian noise with standard deviation σ. The proximity operator of $\frac{\lambda}{2}(\cdot - v_m)^2$ is given by $\mathrm{prox}_{\frac{\lambda}{2}(\cdot - v_m)^2}(x) = \frac{\lambda v_m + x}{1+\lambda}$.

For a test image, we used a color *Castle* image taken from [36] of size 256×256 ($N = 256^2 \times 3$). The blur matrix $\mathbf{\Phi}$ was made from spatially-varying (per-pixel) kernels simulating motion-blur. Since the pixels of a blurred image $\mathbf{v}$ are spatially correlated to each other, we tested the two ways of mini-batch construction: (i) simple block partitioning and (ii) the spatially-uniform sampling proposed in Sec. 3.2. The noise standard deviation is set to $\sigma = 2.55$, and the parameter of the data-fidelity is chosen as $\lambda = 1000$.

The right of Fig. 1 plots the convergence behavior of PDS, L-ADMM, and Alg. 2 on the non-uniform deblurring experiment, where the three different numbers of mini-batches: $b = 4, 16, 64$ are examined for Alg. 2 (for a simple implementation of (ii), we set the number of mini-bathes to be squared numbers). One sees that the proposed method is not much more efficient, even slower in some cases, than PDS and L-ADMM, which is different from the case of the CS reconstruction experiment. This is because the blur matrix $\mathbf{\Phi}$ is relatively sparse, so that the computational advantage of mini-batch decomposition becomes small compared with the case of the dense CS measurement matrix. Hence in such cases, the number of mini-batches should be reasonably small (but not too small not to spoil the benefit of stochastic optimization). Indeed, the proposed method still outperforms PDS and L-ADMM with $b = 4$ and 16. We also remark that the use of our mini-batch construction strategy (ii) results in much faster convergence than the use of the trivial way (i), which demonstrates that the proposed strategy is effective for spatially correlated cases. Finally, as in the CS reconstruction experiment, we observe that the deblurred images in Fig. 2 (bottom) indicate the same PSNR (26.12 [dB]).

5. CONCLUDING REMARKS

We have proposed an efficient image restoration framework based on stochastic proximal optimization. Since the proposed method does not require the multiplication of $\mathbf{\Phi}$ and variables at each iteration, it would be a powerful choice when the structure of $\mathbf{\Phi}$ is complicated.

Although we focus on convex optimization situations, the proposed method can be applied to image restoration with nonconvex objectives, if the proximity(-like) operator of each function is computable, e.g., the ℓ_0 pseudo-norm. With slight modification, one can also use it for image restoration with separated components, such as a recently proposed cartoon-texture decomposition [37].

6. REFERENCES

[1] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, no. 1-4, pp. 259–268, 1992.

[2] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds., pp. 185–212. Springer-Verlag, New York, 2011.

[3] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithm for ℓ_1 minimization with applications to compressed sensing," *SIAM J. Imag. Sci.*, vol. 1, no. 1, pp. 143–168, 2008.

[4] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. (NIPS)*, 2011.

[5] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2010.

[6] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optimization Theory and Applications*, 2013.

[7] P. L. Combettes and J.-C. Pesquet, "Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators," *Set-Valued and Variational Analysis*, vol. 20, no. 2, pp. 307–330, 2012.

[8] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Scholkopf, "Fast removal of non-uniform camera shake," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011.

[9] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform deblurring for shaken images," *Int. J. Comput. Vis.*, vol. 98, no. 2, pp. 168–186, 2012.

[10] L. Xu, S. Zheng, and J. Jia, "Unnatural L0 sparse representation for natural image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013.

[11] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Process. Magazine*, vol. 24, no. 4, 2007.

[12] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[13] T. Suzuki, "Stochastic dual coordinate ascent with alternating direction method of multipliers," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014.

[14] S. Ono, T. Miyata, and I. Kumazawa, "Image restoration by stochastic proximal optimization," Tech. Rep., IEICE, Mar. 2015.

[15] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *C. R. Acad. Sci. Paris Ser. A Math.*, vol. 255, pp. 2897–2899, 1962.

[16] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, Springer, New York, 2011.

[17] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Proc. Adv. Neural Inf. Process. (NIPS)*, 2014, pp. 1646–1654.

[18] A. Repetti, E. Chouzenoux, and J.-C. Pesquet, "A random block-coordinate primal-dual proximal algorithm with application to 3D mesh denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014.

[19] J.-C. Pesquet and A. Repetti, "A class of randomized primal-dual algorithms for distributed optimization," *arXiv preprint arXiv:1406.6404*, 2014.

[20] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, and S. McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *arXiv preprint arXiv:1505.00273*, 2015.

[21] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Sel. Topics in Signal Process.*, vol. 1, pp. 564–574, 2007.

[22] X. Bresson and T. F. Chan, "Fast dual minimization of the vectorial total variation norm and applications to color image processing," *Inverse Probl. Imag.*, vol. 2, no. 4, pp. 455–484, 2008.

[23] B. Goldluecke, E. Strekalovskiy, and D. Cremers, "The natural vectorial total variation which arises from geometric measure theory," *SIAM J. Imag. Sci.*, vol. 5, no. 2, pp. 537–563, 2012.

[24] S. Ono and I. Yamada, "Decorrelated vectorial total variation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014.

[25] E. Candès, L. Demanet, D. L. Donoho, and L. Ying, "Fast discrete curvelet transforms," *SIAM J. Multi. Model. Simul.*, vol. 5, no. 3, pp. 861–899, 2006.

[26] G. Gilboa and S. Osher, "Nonlocal linear image regularization and supervised segmentation," *Multiscale Model. Simul.*, vol. 6, no. 2, pp. 595–630, 2007.

[27] A. Danielyan, V. Katkovnik, and K. Egiazarian, "BM3D frames and variational image deblurring," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1715–1728, 2012.

[28] G. Chierchia, N. Pustelnik, B. Pesquet-Popescu, and J.-C. Pesquet, "A nonlocal structure tensor-based approach for multicomponent image recovery problems," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5531–5544, 2014.

[29] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2009.

[30] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011.

[31] H. Ouyang, N. He, L. Tran, and A. Gray, "Stochastic alternating direction method of multipliers," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013.

[32] N. Pustelnik, C. Chaux, and J.-C. Pesquet, "Parallel proximal algorithm for image restoration using hybrid regularization," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2450–2462, 2011.

[33] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[34] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[35] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, 2007.

[36] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2001.

[37] S. Ono, T. Miyata, and I. Yamada, "Cartoon-texture image decomposition using blockwise low-rank texture characterization," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1128–1142, 2014.